CARM: Cache-Aware Roofline Modeling

Performance, Power and Energy-Efficiency

Aleksandar Ilic,

Frederico Pratas and Leonel Sousa

Instituto Superior Técnico, INESC-ID

Universidade de Lisboa, Portugal

ENS Cachan

Université Paris-Saclay, France

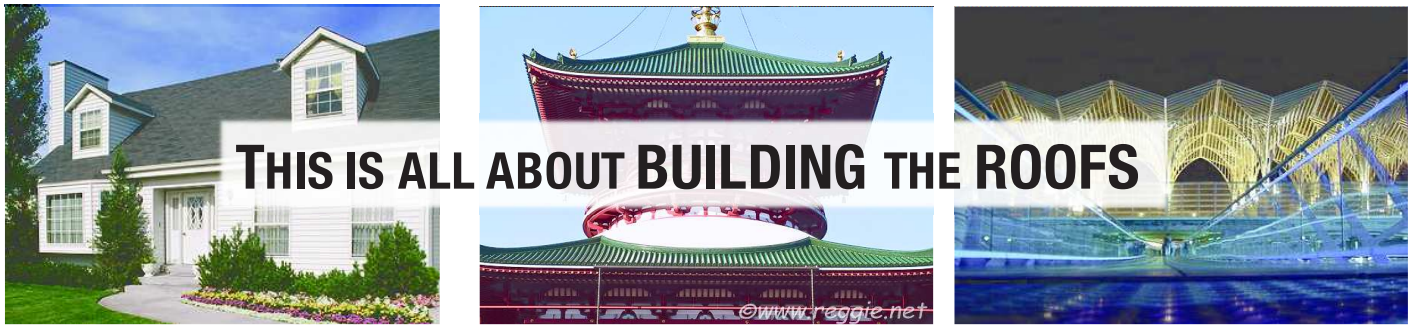
Seminaire Exceptionnel, ENS Cachan, France

November 30, 2015

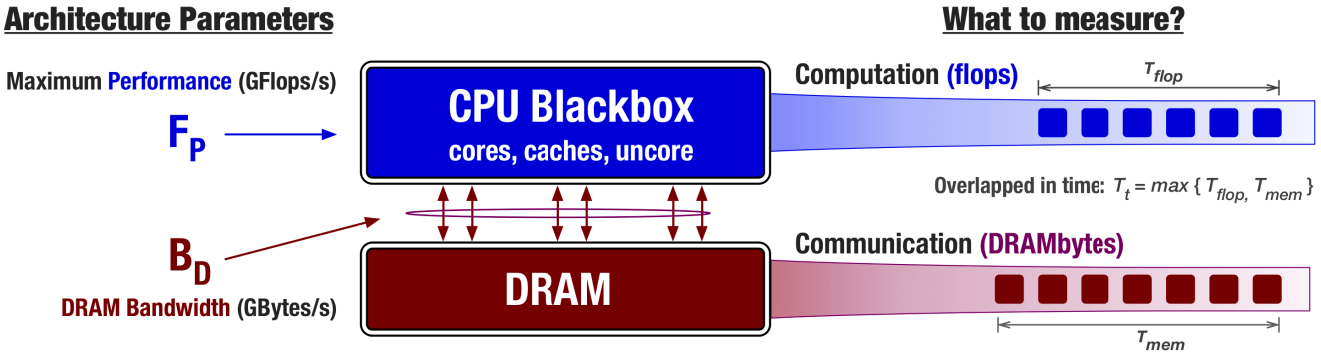
Introduction

Motivation

- **Energy-efficient multi-cores:** Sprinting towards Dark Silicon solutions
  - **Heterogeneity**: compute and memory capability, specialization (CPU, GPU, big.LITTLE, FPGA)
  - Performance, power, energy and efficiency **trade-offs**
- **Application design**: Exploiting architecture diversity to reach efficiency
  - Different **computational and memory requirements**
  - **Painful** optimization and characterization (for each architecture)
- How far can we go: **Performance, Power, Energy and Efficiency Maximums?**
  - In a simple, **insightful** and fast way (allowing the first-order analysis)



State of the art: Original Roofline Model

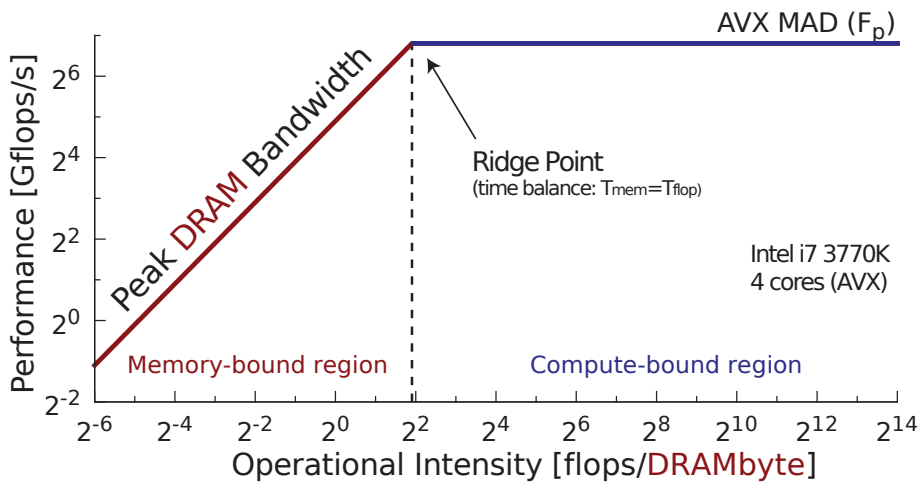


Original Roofline Model*:

Attainable Performance (Gflops/s)

$F_a = flops/T_t = min\{F_p, B_D \times I\}$

Operational Intensity (flops/DRAMbytes)

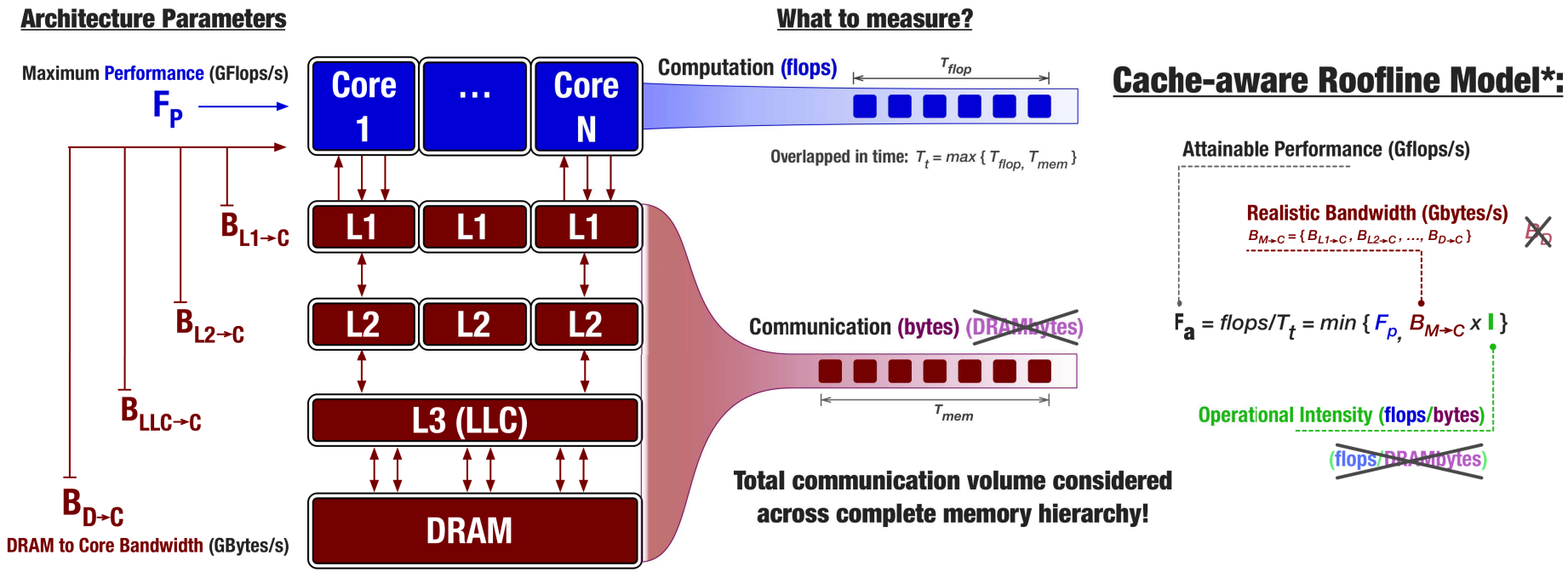


Known Limitations**:

- Targets **single-memory-level architectures**
  - In practice, **several models** for each hierarchy level
- **Distorted characterization** of real applications
  - **Ignored** "bytes" from **caches** cause **shift in *I***
  - Memory-bound application can become compute-bound
- Limited **practical validation**: Time paradox
  - Hitting memory lines requires specific clock counters
- Mostly pedagogical

** Williams S., et. al., "Roofline: An insightful visual performance model for multicore architectures", Communications of the ACM (2009)*
*** Ilic A., Pratas F., Sousa L., "Cache-aware Roofline Model: Upgrading the loft", IEEE CAL (2014)*

Proposed Work: Cache-aware Roofline Model (CARM)

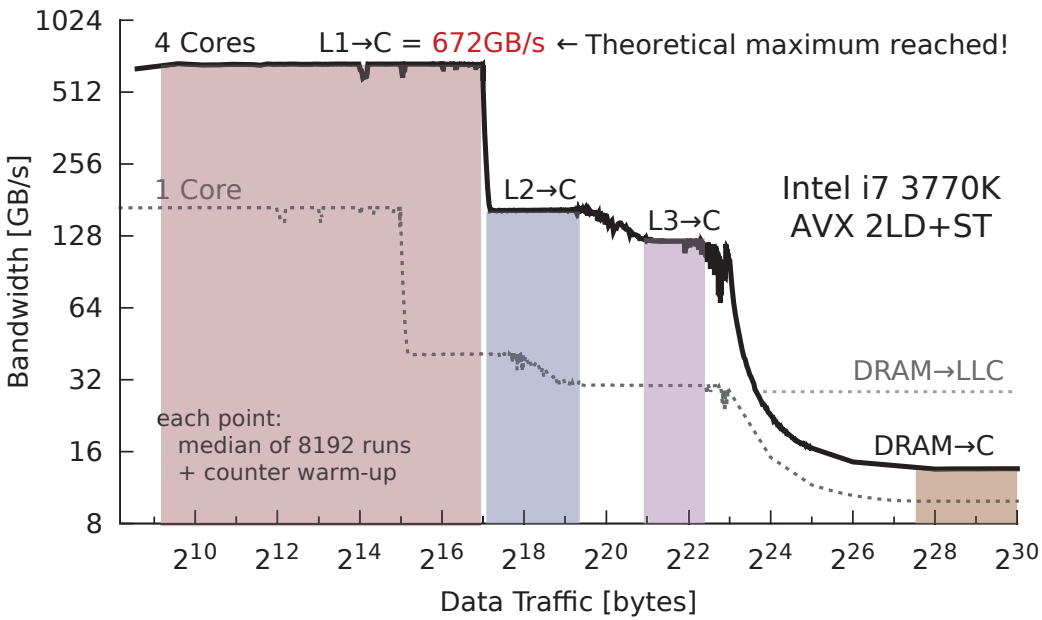


- CARM considers memory traffic and computation from the **consistent architecture point of view** (cores)
  - **No time paradox**: Possible **practical validation** and all **modeled maximums can be hit** in reality
  - **No distorted perspective** when optimizing applications (**no shift in I** - traffic across all memory levels considered)
  - Unifies the complete memory hierarchy in a **single plot model**
- **PROBLEM**: How to obtain these bandwidth values? (only $B_{L1\text{->}C}$ can be directly derived from data sheets)

** Ilic A., Pratas F., Sousa L., "Cache-aware Roofline Model: Upgrading the loft", IEEE CAL (2014)*

Developed Tools: Methodology and micro-benchmarks

- **CARM-oriented tools** for precise monitoring based on hardware counters
  - **KerMon**[1]: Integration of hardware counters inside the Linux scheduler (patching required)
  - **SchedMon**[2]: Software tool for near-OS counter-based monitoring (multiplexing and shed events)
  - **Custom** validation tools and micro-benchmarks[3]

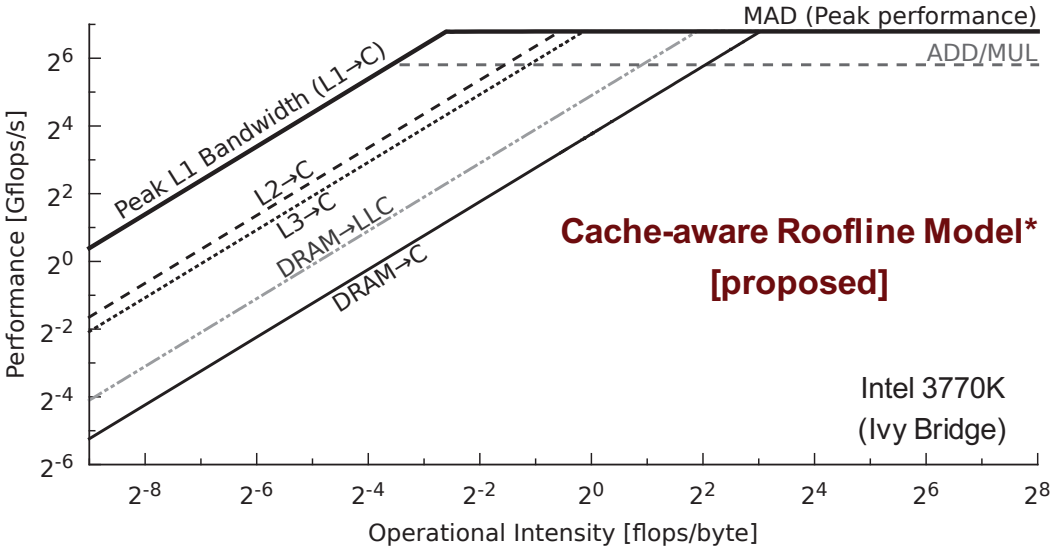



```
// Configured Counters
CPU_CLK_UNHALTED.CORE/REF
MEM_UOP_RETIRED.ALL_LOADS
MEM_UOP_RETIRED.ALL_STORES
…
```

```
// AVX Assembly: 2LD+1ST
vmovapd   0(%rax), %ymm0
vmovapd  32(%rax), %ymm1
vmovapd      %ymm2, 64(%rax)
vmovapd  96(%rax), %ymm3
vmovapd 128(%rax), %ymm4
vmovapd      %ymm5, 160(%rax)
…
```

[1] *Antão D., et. al. "Monitoring Performance and Power for Application Characterization with Cache-aware Roofline Model", PPAM (2013)*
[2] *Taniça L., et. al. "SchedMon: A Performance and Energy Monitoring Tool for Modern Multi-cores", MuCoCoS/Euro-Par (2014)*
[3] *Ilic A., "Heterogeneous Systems: Load Balancing and Performance Modeling", PhD Thesis, IST, Lisbon (2014)*

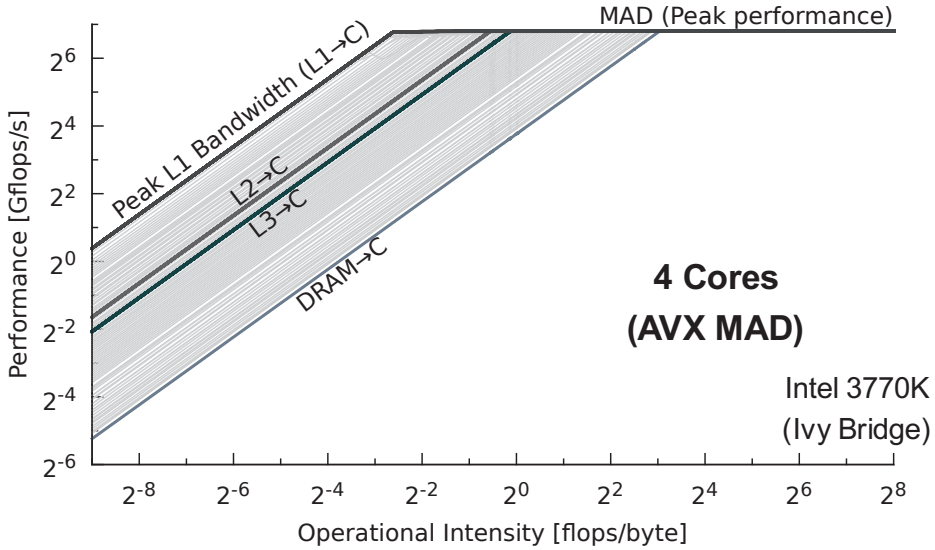
Cache-aware Roofline Model: Performance



- Insightful **single plot model**
  - Shows performance limits of multicores
  - Redefined OI: flops and bytes as seen by core
  - Constructed once per architecture
- Considers **complete memory hierarchy**
  - Influence of caches and DRAM to performance
- Applicable to **other types of operations**
  - not only floating-point
- **Useful for:**
  - **Application** characterization and optimization
  - **Architecture** development and understanding

** Ilic A., Pratas F., Sousa L., "Cache-aware Roofline Model: Upgrading the loft", IEEE CAL (2014)*

Cache-aware Roofline Model: Performance

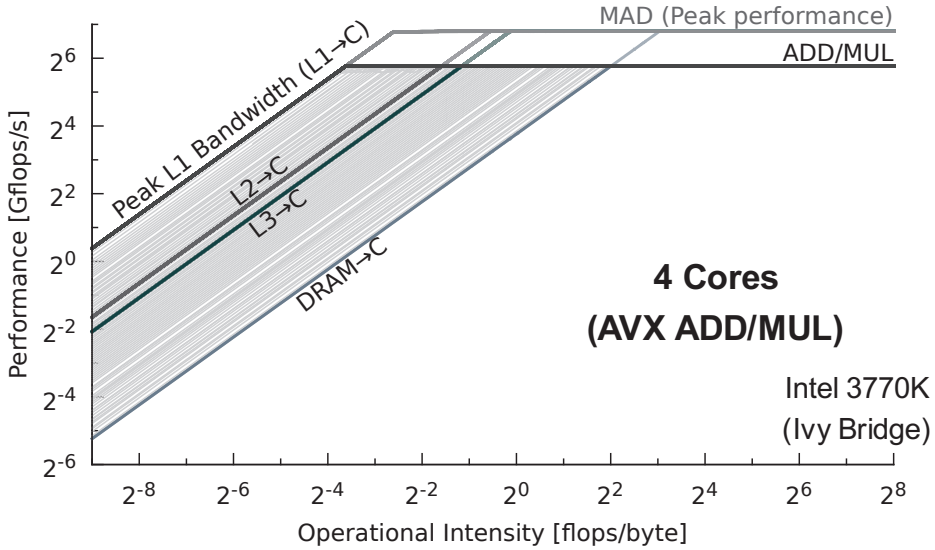


- **Total Cache-aware Roofline Model**
  - Includes **all transitional states** (traversing the memory hierarchy and filling the pipeline)
  - Single-plot modeling for **different** types of compute and memory **operations**

- Insightful **single plot model**
  - Shows performance limits of multicores
  - Redefined OI: flops and bytes as seen by core
  - Constructed once per architecture
- Considers **complete memory hierarchy**
  - Influence of caches and DRAM to performance
- Applicable to **other types of operations**
  - not only floating-point
- **Useful for:**
  - **Application** characterization and optimization
  - **Architecture** development and understanding

** Ilic A., Pratas F., Sousa L., "Cache-aware Roofline Model: Upgrading the loft", IEEE CAL (2014)*

Cache-aware Roofline Model: Performance

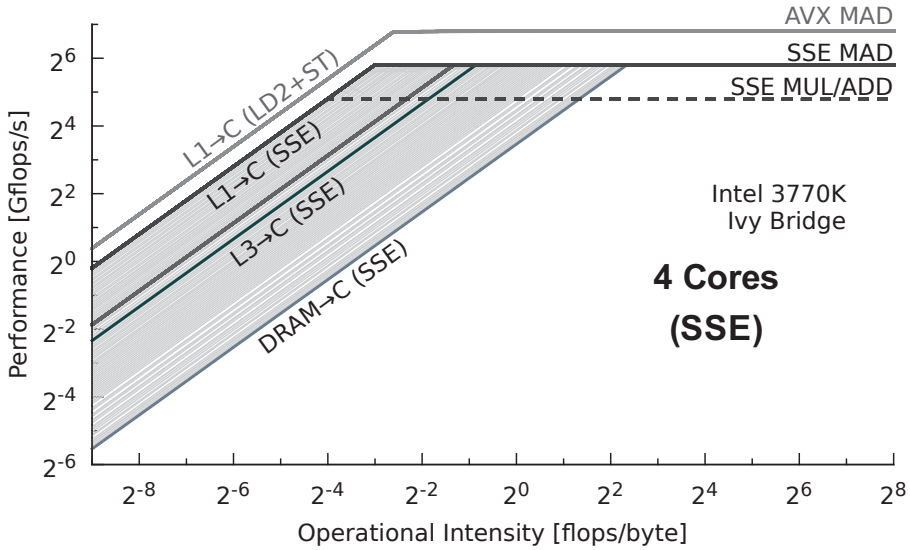


- **Total Cache-aware Roofline Model**
  - Includes **all transitional states** (traversing the memory hierarchy and filling the pipeline)
  - Single-plot modeling for **different** types of compute and memory **operations**

- Insightful **single plot model**
  - Shows performance limits of multicores
  - Redefined OI: flops and bytes as seen by core
  - Constructed once per architecture
- Considers **complete memory hierarchy**
  - Influence of caches and DRAM to performance
- Applicable to **other types of operations**
  - not only floating-point
- **Useful for:**
  - **Application** characterization and optimization
  - **Architecture** development and understanding

** Ilic A., Pratas F., Sousa L., "Cache-aware Roofline Model: Upgrading the loft", IEEE CAL (2014)*

Cache-aware Roofline Model: Performance

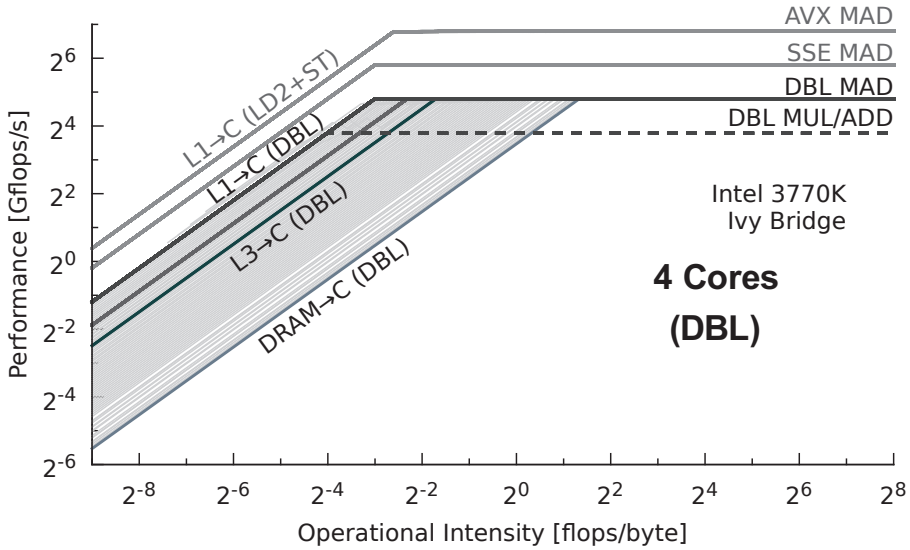


- **Total Cache-aware Roofline Model**
  - Includes **all transitional states** (traversing the memory hierarchy and filling the pipeline)
  - Single-plot modeling for **different** types of compute and memory **operations**

- Insightful **single plot model**
  - Shows performance limits of multicores
  - Redefined OI: flops and bytes as seen by core
  - Constructed once per architecture
- Considers **complete memory hierarchy**
  - Influence of caches and DRAM to performance
- Applicable to **other types of operations**
  - not only floating-point
- **Useful for:**
  - **Application** characterization and optimization
  - **Architecture** development and understanding

** Ilic A., Pratas F., Sousa L., "Cache-aware Roofline Model: Upgrading the loft", IEEE CAL (2014)*

Cache-aware Roofline Model: Performance



- **Total Cache-aware Roofline Model**
  - Includes **all transitional states** (traversing the memory hierarchy and filling the pipeline)
  - Single-plot modeling for **different** types of compute and memory **operations**

- Insightful **single plot model**
  - Shows performance limits of multicores
  - Redefined OI: flops and bytes as seen by core
  - Constructed once per architecture
- Considers **complete memory hierarchy**
  - Influence of caches and DRAM to performance
- Applicable to **other types of operations**
  - not only floating-point
- **Useful for:**
  - **Application** characterization and optimization
  - **Architecture** development and understanding

** Ilic A., Pratas F., Sousa L., "Cache-aware Roofline Model: Upgrading the loft", IEEE CAL (2014)*

Outline

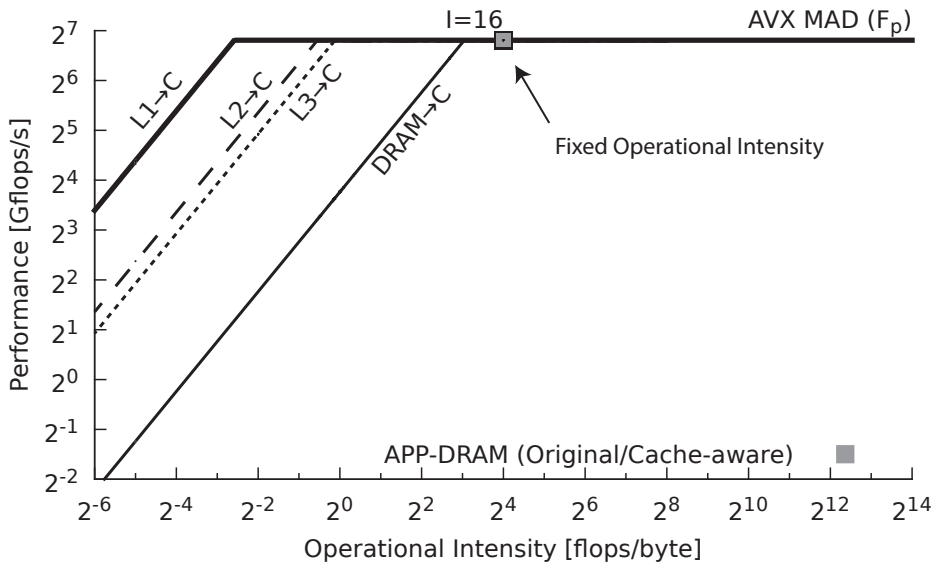
CACHE-AWARE ROOFLINE MODEL

- PERFORMANCE
- POWER CONSUMPTION
- EFFICIENCY

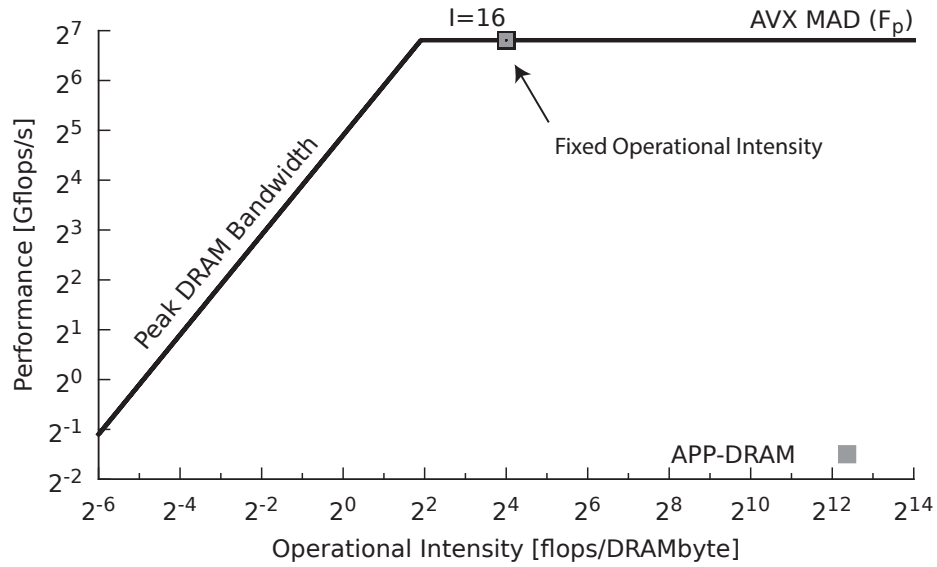
Hands-on:
Application Characterization

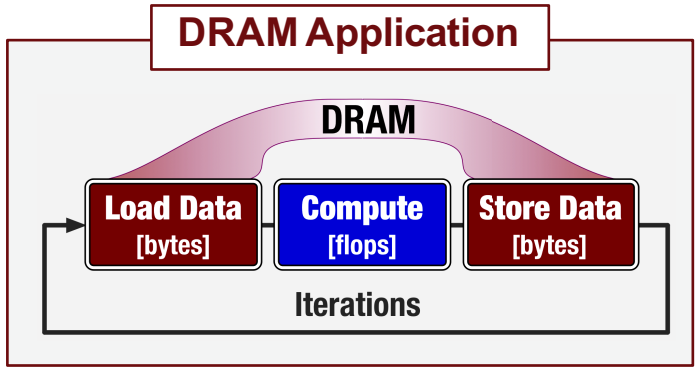
Application Characterization: DRAM

Cache-aware Roofline Model (CARM)[1]



Original Roofline Model (ORM)[2]

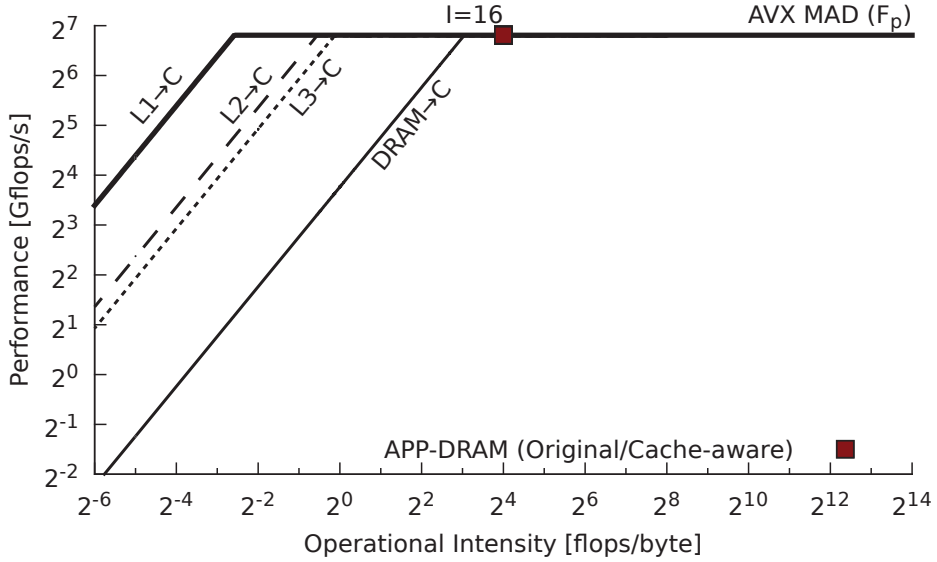




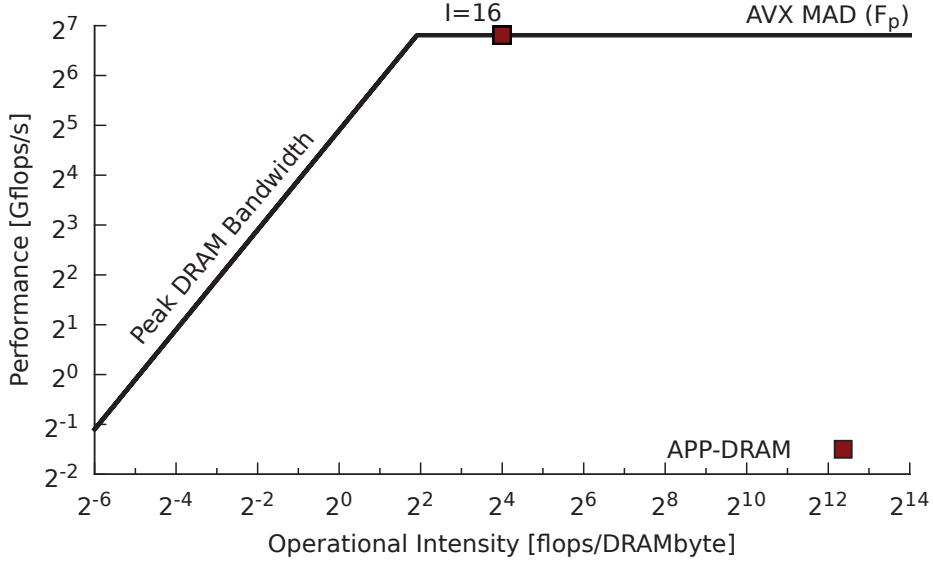
1) Ilic A., Pratas F., Sousa L., "Cache-aware Roofline Model: Upgrading the loft", IEEE CAL (2014)
2) Williams S., et. al., "Roofline: An insightful visual performance model for multicore architectures", Communications of the ACM (2009)

Application Characterization: DRAM

Cache-aware Roofline Model (CARM)[1]



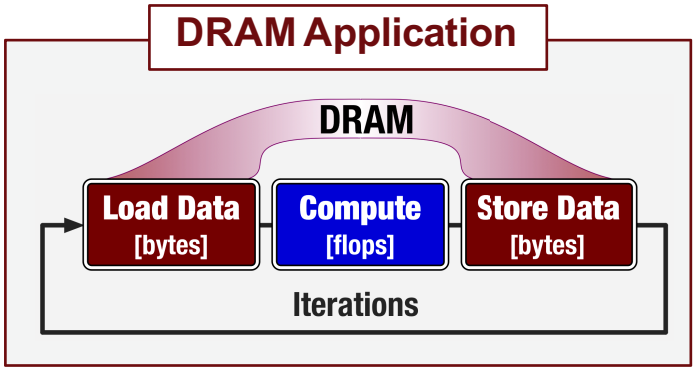
Original Roofline Model (ORM)[2]



Cache-aware Roofline Model:

iterations:	1	2	3	...	N
flops:	f	f	f	f	f
bytes:	b	b	b	b	b
operational intensity:	$\frac{f}{b}$	$\frac{2f}{2b}$	$\frac{3f}{3b}$	...	$\frac{Nf}{Nb}$

f/b (constant)



Original Roofline Model:

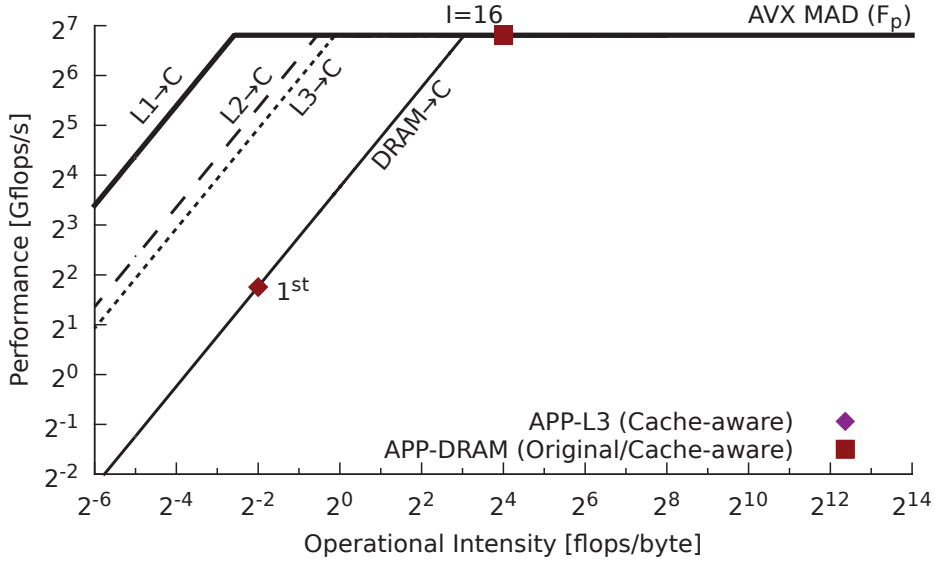
iterations:	1	2	3	...	N
flops:	f	f	f	f	f
DRAMbytes:	b'	b'	b'	b'	b'
operational intensity:	$\frac{f}{b'}$	$\frac{2f}{2b'}$	$\frac{3f}{3b'}$	...	$\frac{Nf}{Nb'}$

f/b' (constant)

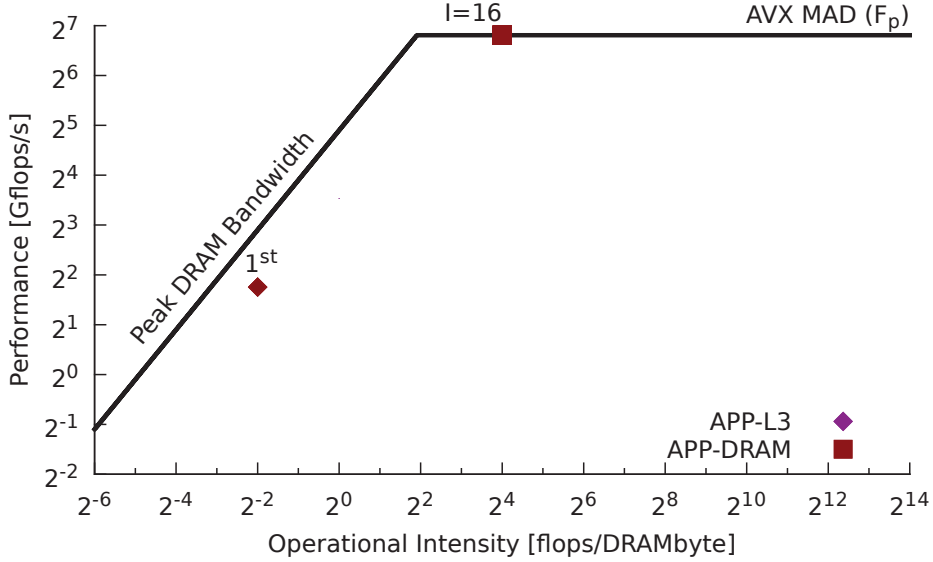
1) Ilic A., Pratas F., Sousa L., "Cache-aware Roofline Model: Upgrading the loft", IEEE CAL (2014)
2) Williams S., et. al., "Roofline: An insightful visual performance model for multicore architectures", Communications of the ACM (2009)

Application Characterization: DRAM

Cache-aware Roofline Model (CARM)[1]

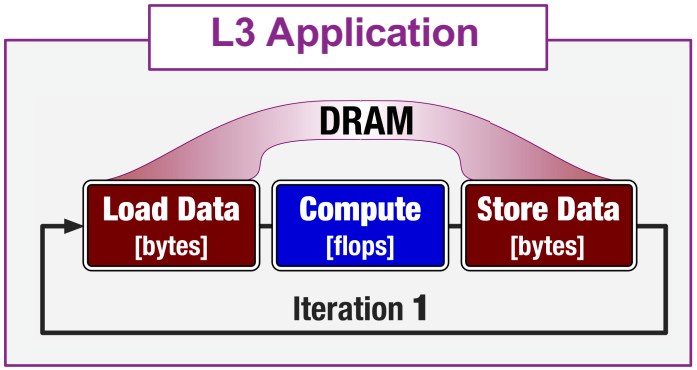


Original Roofline Model (ORM)[2]



Cache-aware Roofline Model:

iterations:	1
flops:	f
bytes:	b
operational intensity:	$\frac{f}{b}$



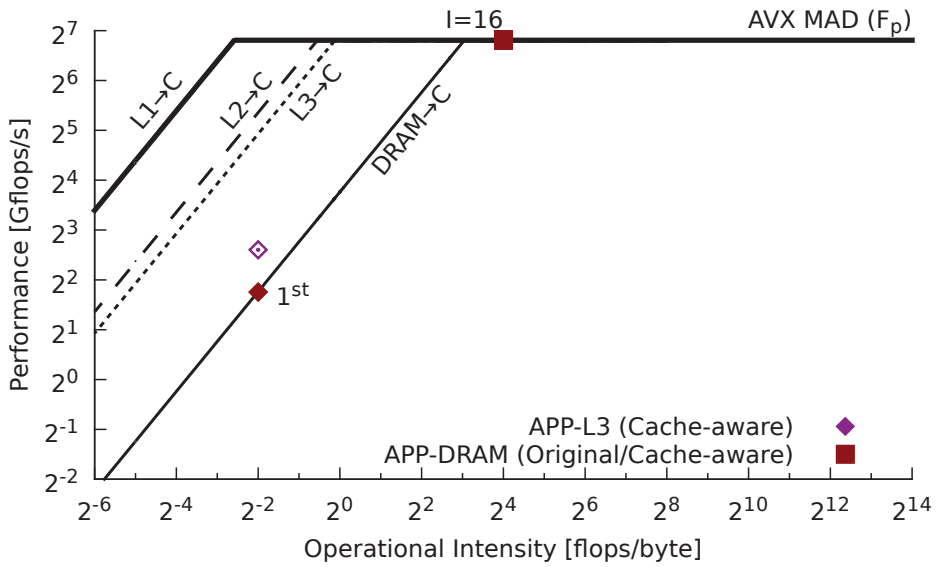
Original Roofline Model:

iterations:	1
flops:	f
DRAMbytes:	b'
operational intensity:	$\frac{f}{b'}$

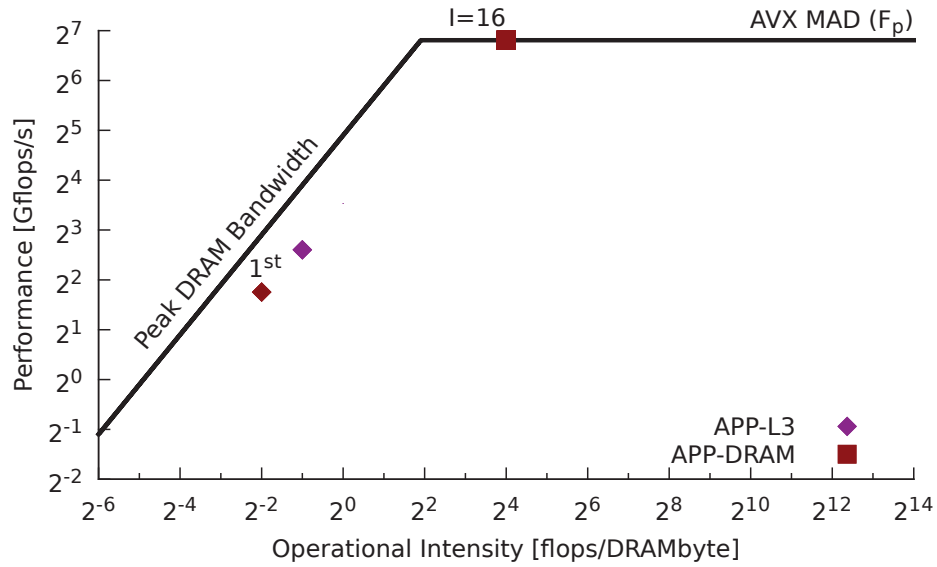
1) Ilic A., Pratas F., Sousa L., "Cache-aware Roofline Model: Upgrading the loft", IEEE CAL (2014)
2) Williams S., et. al., "Roofline: An insightful visual performance model for multicore architectures", Communications of the ACM (2009)

Application Characterization: DRAM

Cache-aware Roofline Model (CARM)[1]

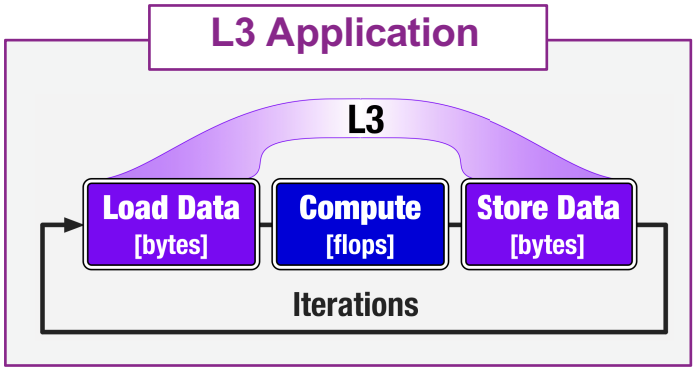


Original Roofline Model (ORM)[2]



Cache-aware Roofline Model:

iterations:	1	2
flops:	f	f
bytes:	b	b
operational intensity:	$\frac{f}{b}$	$\frac{2f}{b+b}$



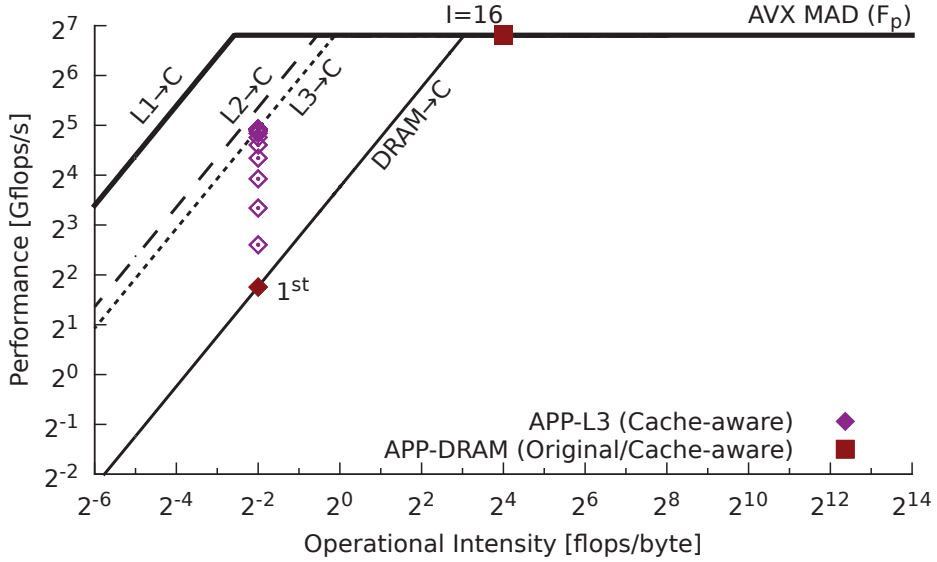
Original Roofline Model:

iterations:	1	2
flops:	f	f
DRAMbytes:	b'	
operational intensity:	$\frac{f}{b'}$	$\frac{2f}{b'}$

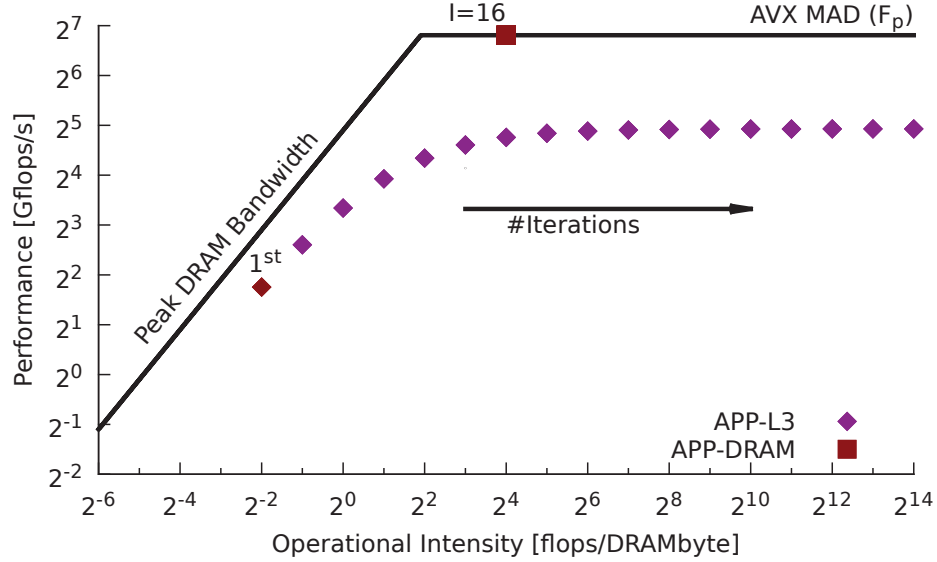
1) Ilic A., Pratas F., Sousa L., "Cache-aware Roofline Model: Upgrading the loft", IEEE CAL (2014)
2) Williams S., et. al., "Roofline: An insightful visual performance model for multicore architectures", Communications of the ACM (2009)

Application Characterization: DRAM

Cache-aware Roofline Model (CARM)[1]



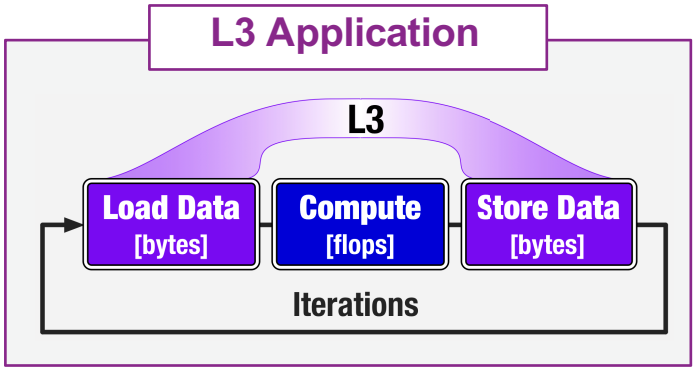
Original Roofline Model (ORM)[2]



Cache-aware Roofline Model:

iterations:	1	2	3	...	N
flops:	f	f	f	f	f
bytes:	b	b	b	b	b
operational intensity:	$\frac{f}{b}$	$\frac{2f}{b+b}$	$\frac{3f}{b+2b}$	...	$\frac{Nf}{b+(N-1)b}$

f/b (constant)



Original Roofline Model:

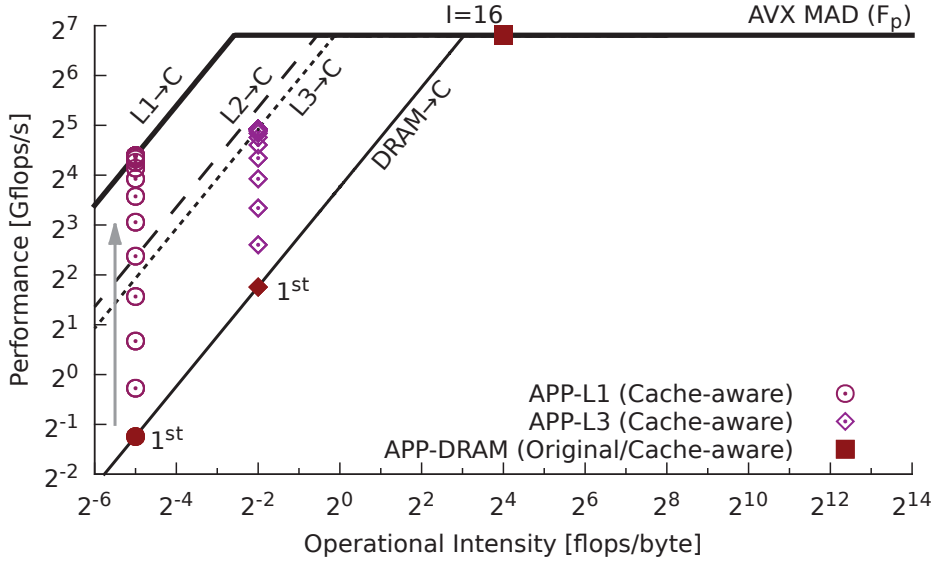
iterations:	1	2	3	...	N
flops:	f	f	f	f	f
DRAMbytes:	b'				
operational intensity:	$\frac{f}{b'}$	$\frac{2f}{b'}$	$\frac{3f}{b'}$	...	$\frac{Nf}{b'}$

??? (shift right)

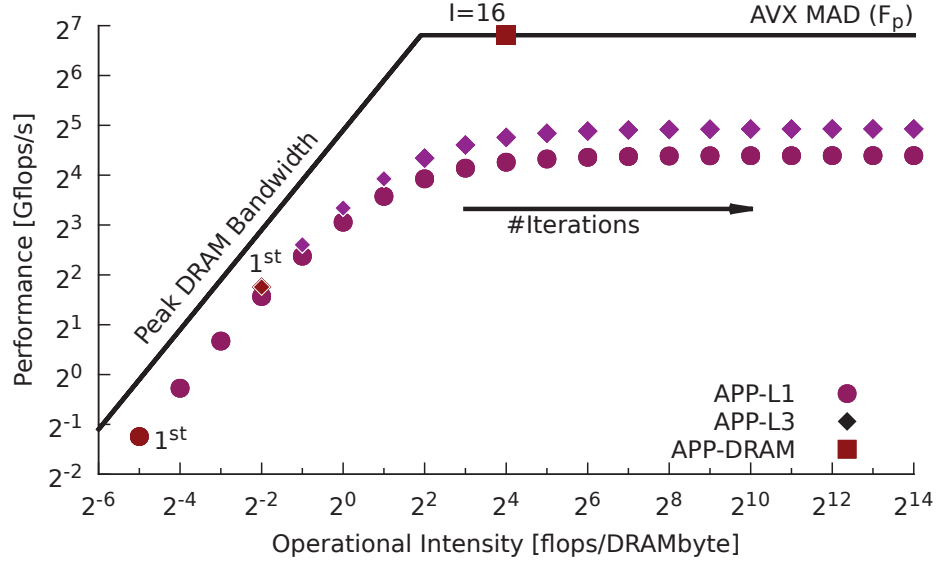
1) Ilic A., Pratas F., Sousa L., "Cache-aware Roofline Model: Upgrading the loft", IEEE CAL (2014)
2) Williams S., et. al., "Roofline: An insightful visual performance model for multicore architectures", Communications of the ACM (2009)

Application Characterization: DRAM

Cache-aware Roofline Model (CARM)[1]



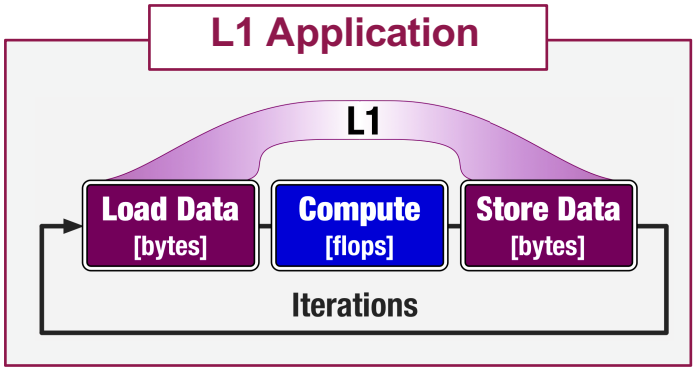
Original Roofline Model (ORM)[2]



Cache-aware Roofline Model:

iterations:	1	2	3	…	N
flops:	f	f	f	f	f
bytes:	b	b	b	b	b
operational intensity:	$\frac{f}{b}$	$\frac{2f}{b+b}$	$\frac{3f}{b+2b}$	…	$\frac{Nf}{b+(N-1)b}$

f/b (constant)



Original Roofline Model:

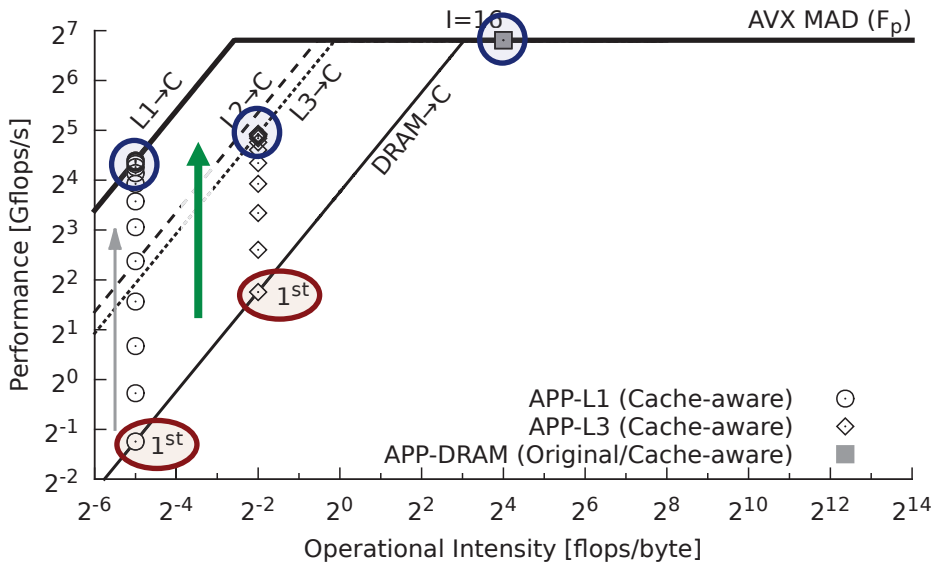
iterations:	1	2	3	…	N
flops:	f	f	f	f	f
DRAMbytes:	b'				
operational intensity:	$\frac{f}{b'}$	$\frac{2f}{b'}$	$\frac{3f}{b'}$	…	$\frac{Nf}{b'}$

??? (shift right)

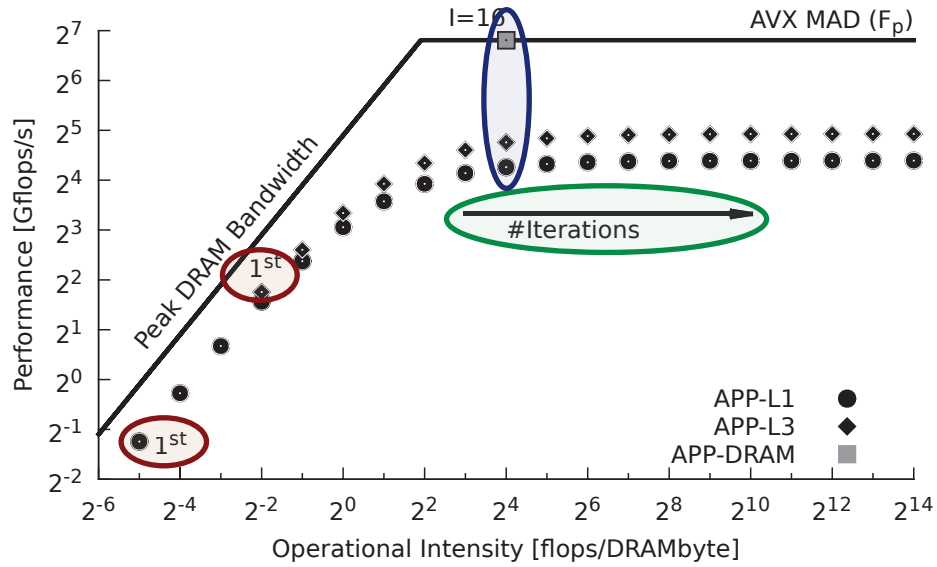
1) Ilic A., Pratas F., Sousa L., "Cache-aware Roofline Model: Upgrading the loft", IEEE CAL (2014)
2) Williams S., et. al., "Roofline: An insightful visual performance model for multicore architectures", Communications of the ACM (2009)

Application Characterization: DRAM

Cache-aware Roofline Model (CARM)[1]



Original Roofline Model (ORM)[2]



Observations (Paradoxes)	Cache-aware Roofline Model (CARM)	Original Roofline Model (ORM)
Memory-bound Region: Modeled max. performance	**Achievable** (in practice)	**Not achievable** (for architectures with caches)
Operational Intensity	**Constant** (no variation with the problem size)	**Varies with the problem size** (shift from memory-bound to compute-bound)
Characterization: Optimization Hints	**Consistent** (according to the test nature)	**Inconclusive / Unexpected** *Hint*: Optimize memory-bound APP-L1? APP-DRAM performs better than APP-L1

Outline

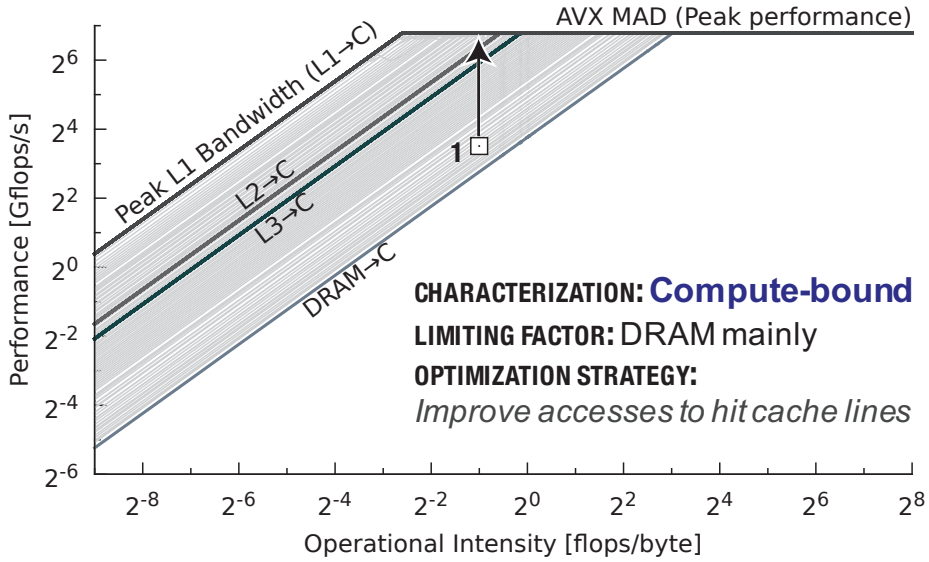
CACHE-AWARE ROOFLINE MODEL

- PERFORMANCE
- POWER CONSUMPTION
- EFFICIENCY

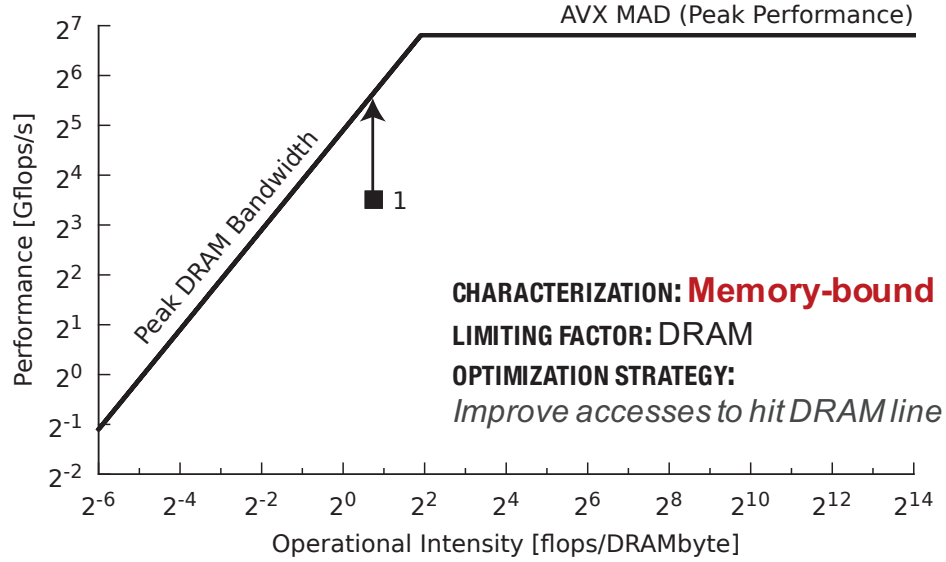
Hands-on:
Application Optimization

Application Optimization: Dense Matrix Multiplication

Cache-aware Roofline Model (CARM)[1]



Original Roofline Model (ORM)[2]

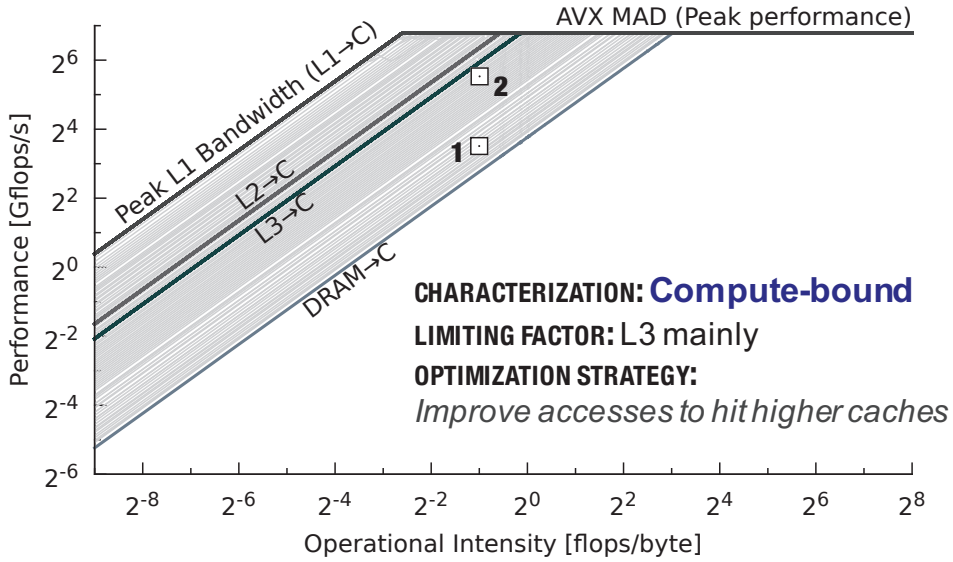


VERSION	OPTIMIZATION STRATEGY
1	**Basic** implementation: Row-major matrices
2	Improved memory access by **transposing B matrix**
3, 4, 5	Blocking for **caches**: **L3 (pt. 3), L2 (pt. 4) and L1 (pt. 5)**
6	Highly optimized **Intel MKL** implementation

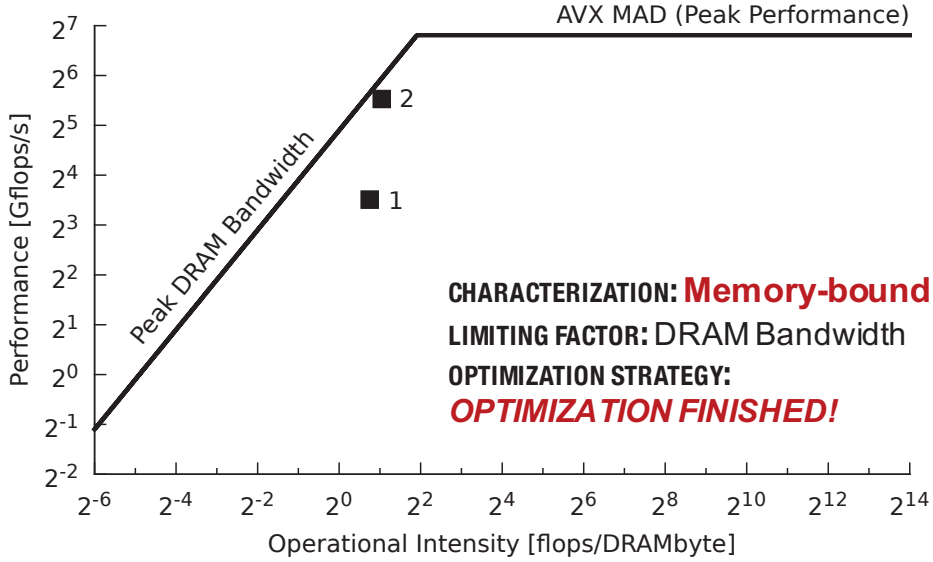
1) *Ilic A., Pratas F., Sousa L., "Cache-aware Roofline Model: Upgrading the loft", IEEE CAL (2014)*
2) *Williams S., et. al., "Roofline: An insightful visual performance model for multicore architectures", Communications of the ACM (2009)*

Application Optimization: Dense Matrix Multiplication

Cache-aware Roofline Model (CARM)[1]



Original Roofline Model (ORM)[2]

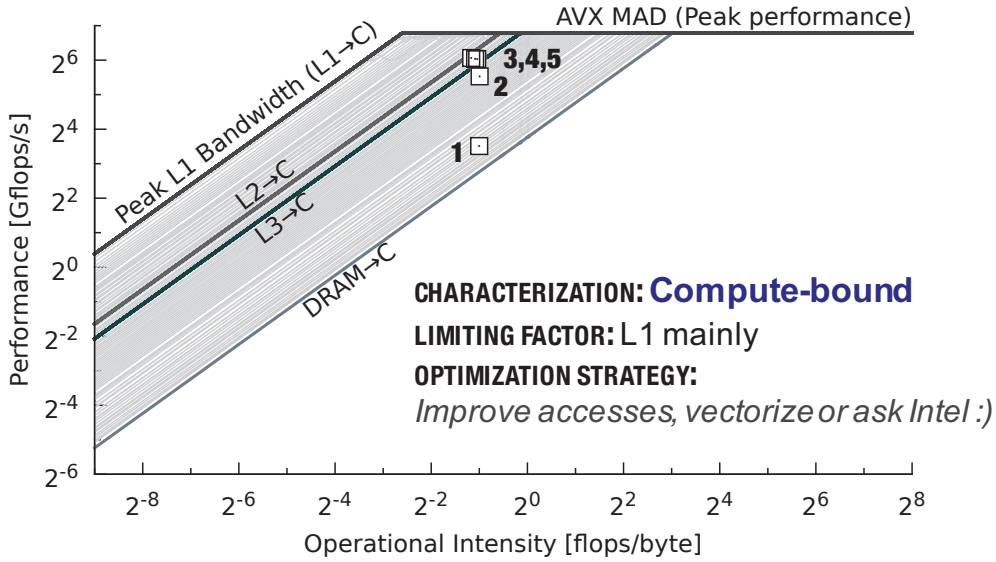


VERSION	OPTIMIZATION STRATEGY
1	**Basic** implementation: Row-major matrices
2	Improved memory access by **transposing B matrix**
3, 4, 5	Blocking for **caches**: **L3 (pt. 3), L2 (pt. 4) and L1 (pt. 5)**
6	Highly optimized **Intel MKL** implementation

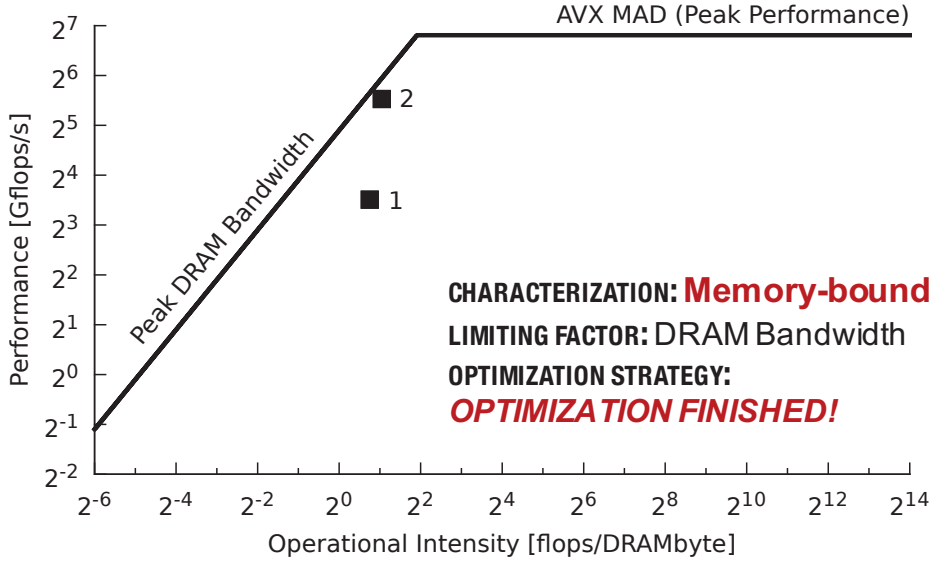
1) *Ilic A., Pratas F., Sousa L., "Cache-aware Roofline Model: Upgrading the loft", IEEE CAL (2014)*
2) *Williams S., et. al., "Roofline: An insightful visual performance model for multicore architectures", Communications of the ACM (2009)*

Application Optimization: Dense Matrix Multiplication

Cache-aware Roofline Model (CARM)[1]



Original Roofline Model (ORM)[2]

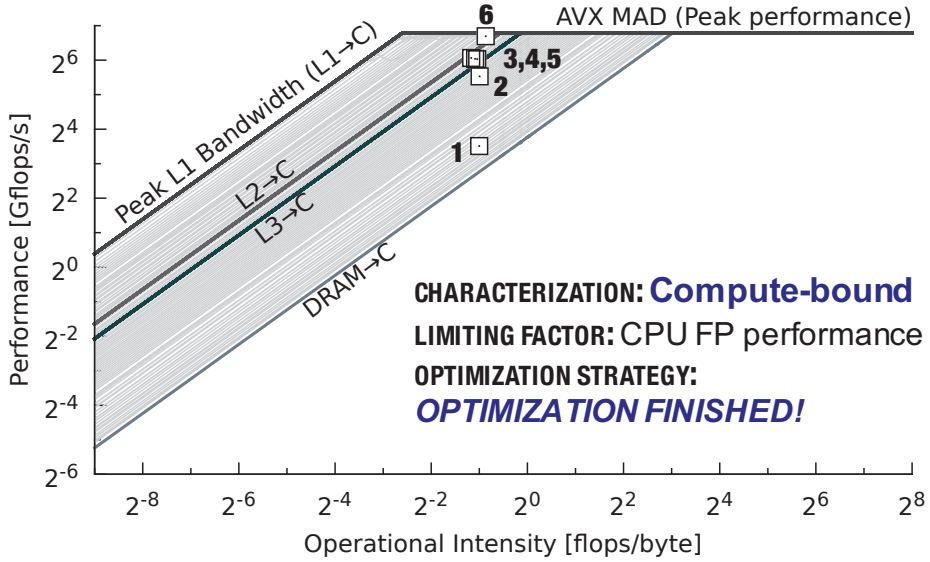


VERSION	OPTIMIZATION STRATEGY
1	**Basic** implementation: Row-major matrices
2	Improved memory access by **transposing B matrix**
3, 4, 5	Blocking for **caches**: **L3 (pt. 3), L2 (pt. 4) and L1 (pt. 5)**
6	Highly optimized **Intel MKL** implementation

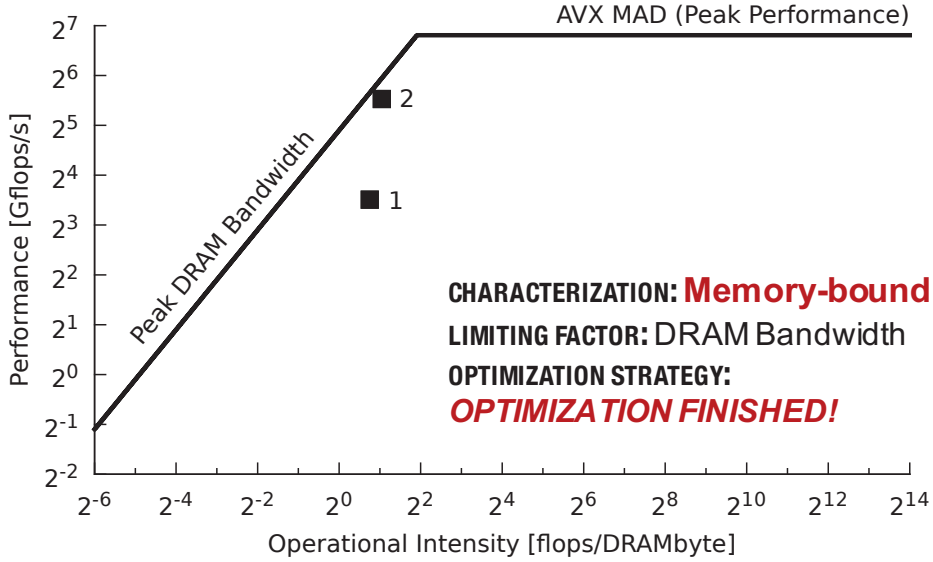
1) *Ilic A., Pratas F., Sousa L., "Cache-aware Roofline Model: Upgrading the loft", IEEE CAL (2014)*
2) *Williams S., et. al., "Roofline: An insightful visual performance model for multicore architectures", Communications of the ACM (2009)*

Application Optimization: Dense Matrix Multiplication

Cache-aware Roofline Model (CARM)[1]



Original Roofline Model (ORM)[2]

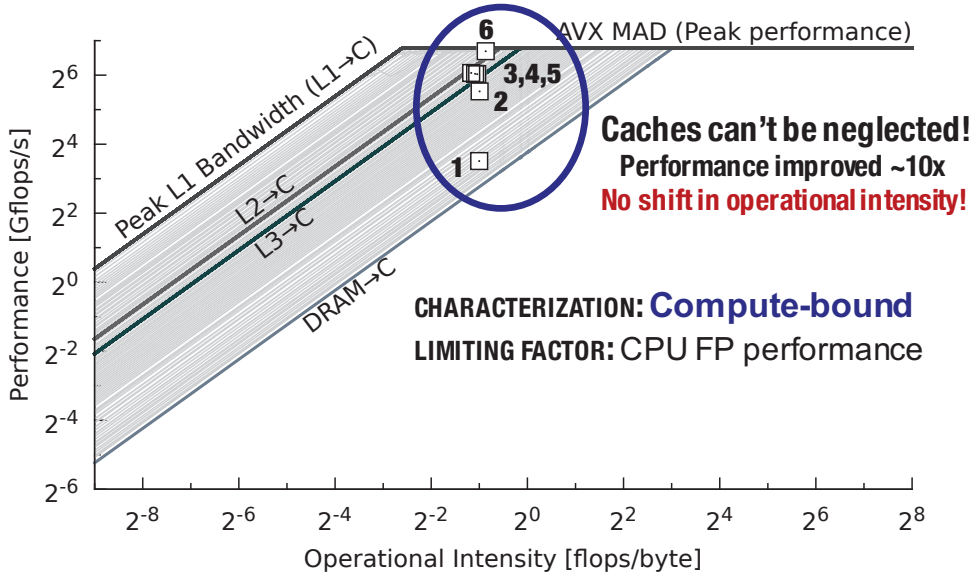


VERSION	OPTIMIZATION STRATEGY
1	**Basic** implementation: Row-major matrices
2	Improved memory access by **transposing B matrix**
3, 4, 5	Blocking for **caches**: **L3 (pt. 3), L2 (pt. 4) and L1 (pt. 5)**
6	Highly optimized **Intel MKL** implementation

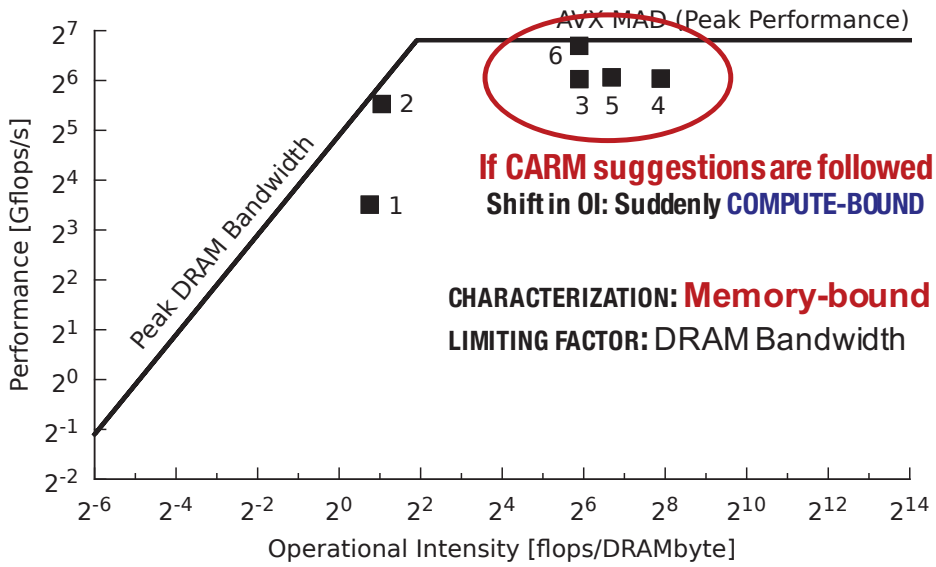
1) *Ilic A., Pratas F., Sousa L., "Cache-aware Roofline Model: Upgrading the loft", IEEE CAL (2014)*
2) *Williams S., et. al., "Roofline: An insightful visual performance model for multicore architectures", Communications of the ACM (2009)*

Application Optimization: Dense Matrix Multiplication

Cache-aware Roofline Model (CARM)[1]



Original Roofline Model (ORM)[2]

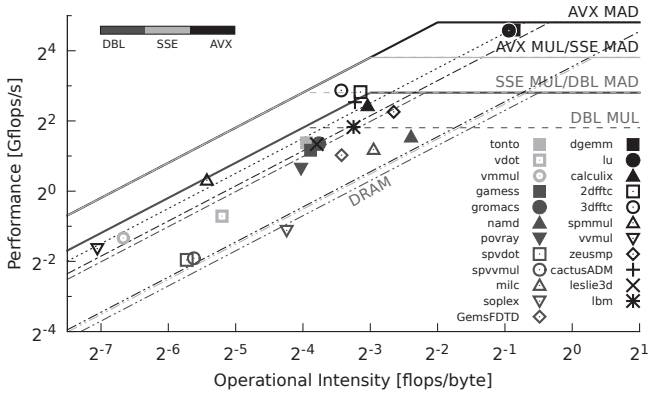


VERSION	OPTIMIZATION STRATEGY
1	**Basic** implementation: Row-major matrices
2	Improved memory access by **transposing B matrix**
3, 4, 5	Blocking for **caches**: **L3 (pt. 3), L2 (pt. 4) and L1 (pt. 5)**
6	Highly optimized **Intel MKL** implementation

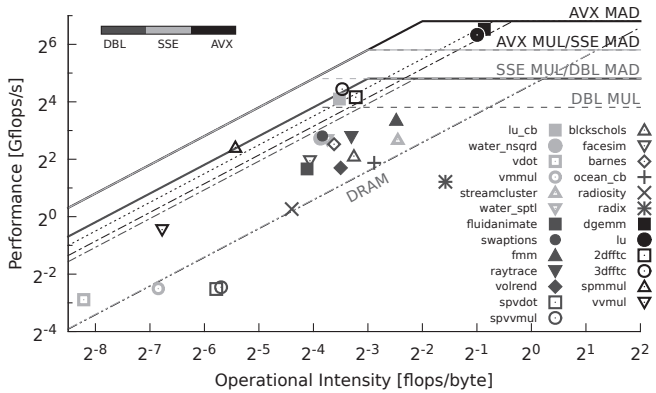
1) *Ilic A., Pratas F., Sousa L., "Cache-aware Roofline Model: Upgrading the loft", IEEE CAL (2014)*
2) *Williams S., et. al., "Roofline: An insightful visual performance model for multicore architectures", Communications of the ACM (2009)*

Cache-aware Roofline Model: Use Cases

Application Characterization

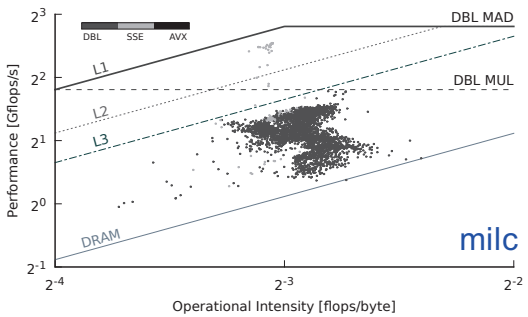


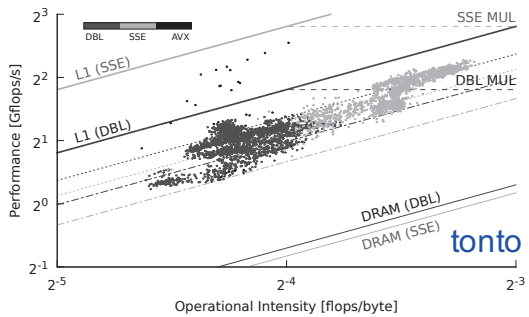
single core: SPECs and MKL

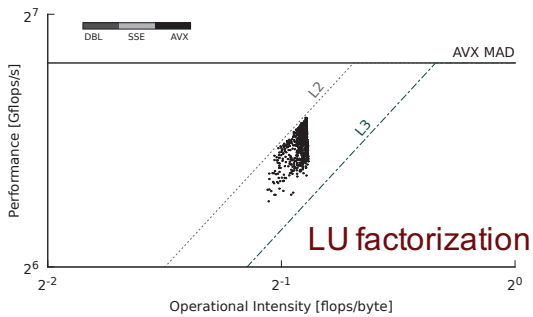


quad-core: PARSEC/SPLASH and MKL

Online Monitoring







1) Antão D., et. al. "Monitoring Performance and Power for Application Characterization with Cache-aware Roofline Model", PPAM (2013)
2) Taniça L., et. al. "SchedMon: A Performance and Energy Monitoring Tool for Modern Multi-cores", MuCoCoS/Euro-Par (2014)
3) Ilic A., "Heterogeneous Systems: Load Balancing and Performance Modeling", PhD Thesis, IST, Lisbon (2014)

Questions?

Thank you!

Further readings:

A. Ilic, F. Pratas, and L. Sousa, ***"Cache-aware Roofline model: Upgrading the loft",*** IEEE Computer Architecture Letters, CAL (2013)

A. Ilic, F. Pratas, and L. Sousa, ***"CARM: Cache-Aware Performance, Power and Energy-Efficiency Roofline Modeling",*** Intel CATC (2015)

L. Taniça, A. Ilic, P. Tomás, and L. Sousa, ***"SchedMon: A Performance and Energy Monitoring Tool for Modern Multi-cores",*** MuCoCoS/Euro-Par (2014)

D. Antão, L. Taniça, A. Ilic, F. Pratas, P. Tomás, and L. Sousa***, "Monitoring Performance and Power for Application Characterization with Cache-aware Roofline Model",*** PPAM (2013)

A. Ilic, F. Pratas, and L. Sousa, ***"Beyond the Roofline: Power, Energy and Efficiency Modeling for Multicores"*** *(#$%&)*